

# Using Microsoft Excel for evaluation data

By Public Health – Seattle & King County, HIV/AIDS Program

---

## Practical tips before you begin

---

- **Naming your variables**

If you're entering data from a survey, it's tempting to use variable names like v1, v2, v3, etc. Instead, use plain language variable names (or abbreviations) so that you don't have to continually refer back to the survey to remember what v27c means. If you are using a survey and want to have the question number in the variable name, then use something like v2race.

To make things easier down the road, don't make the names too long (that just means more typing) and try to keep them to one word without spaces. If you include spaces in the names, you'll end up having to put quotation marks around them every time you use them in a formula.

- **Adding survey numbers**

When entering case records or surveys, always include an ID number. Write it on top of the survey and use it as the first variable in my database. With an ID number, you can always go back to the original survey or record if you find a data error and get the correct information.

- **Save your work often**

There's nothing as frustrating as spending a few hours entering data and then losing it because the program has a fatal error, the computer freezes, or the server crashes. Data entry is not the sort of thing you want to do more of than you must. Also, if you save your worksheet and then do something that totally screws things up and you don't know how to correct the error, you can just close the worksheet (don't save it when Excel asks) and open it up again.

## Navigating around your workbook

---

- **Using worksheets to organize your data**

Rather than entering all your data and doing all your calculations on one worksheet, spread your work on different sheets and then name the sheets. Excel can use information in one sheet to make calculations in another sheet, so I usually put my data in one worksheet and the analyses in other worksheets.

- **Adding worksheets**

When you open Excel, it will show you that you have several sheets to work with. You can add new sheets to accommodate your needs. A workbook can contain no more than 255 sheets. *If you need more than that, you have too much free time.*

Add new sheets by clicking the cursor on the sheet tab at the bottom of the workbook where you want to insert the new sheet, then go to the INSERT menu and click on WORKSHEET.

- **Moving worksheets**

If you want to change the order of the worksheets, click on the tab for the sheet you want to move and drag it to its new position among the other tabs.

- **Naming worksheets**

Rather than leaving the worksheet names as **sheet1**, **sheet2**, etc., rename them so you can remember what they contain. For example, you can name the sheet that contains the data, **data**. If you have a set of tables and charts looking at the effect of race on other variables, name that sheet **race**. You can organize this any way that makes sense to you, just as long as it helps you remember where things are.

To rename a worksheet, double-click on the existing name (Excel will highlight the name) and type in the new name.

- **Using Freeze Panes**

When you have many variables and/or cases, the entire data set will not fit on the screen. Therefore, when you're entering your data, you'll reach a point where you can no longer see your variable names or ID numbers. Rather than try to remember exactly what the rows and columns contain, use the Freeze Panes command to keep the variable names and ID numbers in view as you scroll down or across the worksheet.

To activate Freeze Panes, place the cursor in the cell just to the right of the columns(s) and below the row(s) that you want to always be visible. Go to the Window menu and select Freeze Panes. (This is a toggle switch, so to remove Freeze Panes, go back to the Window menu and select Unfreeze Panes.)

- **Protecting your data**

You've entered your data and now you don't want anyone to mess with it. You can protect the worksheet so people can read it and use it to make calculations, but they can't change any of the entries.

To protect your sheet, place the cursor anywhere in the sheet, go to the Tools menu, click Protection, select Worksheet. Excel will ask you for a password. Beware: If you select a password and then forget what it is (or leave the agency or whatever), no one will be able to access that data). If you don't want to use a password, just click OK and then people can make changes if they unprotect the sheet (same procedure because this is a toggle switch).

We don't use a password because it's usually enough of a reminder to folks when they get the warning that it's protected. You can also protect the entire workbook using the same procedure, but select WORKBOOK instead of WORKSHEET.

## **Cleaning your data**

---

Whenever you enter lots of data into a spreadsheet, you're bound to make some errors. Some errors are easier to catch than others. Say you have a variable with valid codes of 1, 2, and 3. If you look through the spreadsheet and see some 5's and 6's in the column, you know that you have some data entry errors. These are the types of errors that are easy to catch and the section below will help you. Unfortunately, these techniques won't catch those errors where the entry should be 2 and it was entered as 3, which is also a valid code. The only remedy for these errors is to hand check each entry--a tedious task.

- **Using filters**

To activate the autofilter, while in the data sheet, go to the DATA menu, select FILTER, then select AUTOFILTER. When you do this, each variable name will have an arrow by it. Click the arrow and you'll see the values that Excel has detected in the columns. To change invalid codes, click on the code you want to change. Excel will filter out all other codes and the arrow and row numbers will turn blue. For each visible ID number, go back to the original record, find the correct code and change it on the screen. When you're done, click on the filter arrow and select All. Excel will now show all cases.

NOTE: Do this *before* you start making pivot tables (this will make sense when we get there).

- **Using data sort feature**

You can also find (and then correct) errors using the data sort feature. In brief, you sort the data variable by variable, each time looking for invalid values. When you find invalid codes, look at the ID number, go back to the original record, find the actual code, and replace the value in the appropriate cell.

To use data sort, either select a single cell in the dataset or highlight the entire dataset, go to the DATA menu and select SORT. In the dialogue box, select the variable you want to sort by and whether you want to sort in ascending and descending order (for this purpose, it doesn't really matter which you choose), then click OK. You can return the dataset to its original order by sorting by ID number.

**WARNING:** If you highlight just a few columns, Excel will only sort the column(s) you've highlighted and leave the rest in their original order. That, of course, destroys the integrity of the data. Fortunately, the program will give you a warning message and ask you if you really want to go through with your command.

## Naming ranges

---

- **General information**

For those of you who already use formulas, you probably designate the range for the formula as something like A1:A10 or A23:C25. If you have a dataset with 345 cases and 56 variables or if you have data spread over several sheets, that technique becomes cumbersome.

You can already tell that I like naming things and this is no exception. I think it's easier to remember that my HIV status data is in **hivstat** than to remember that it is in F2:F346. It's also easier to remember that my entire dataset is in **database** than to remember that it's in A2:N346. A little work up front will save you a lot of time in the end.

- **Things to do before you begin**

The following preliminary steps allow Excel to recognize the structure of your dataset:

- Make sure that your variable names take only one row of the worksheet.
- Format the variable names differently than the data (e.g. bold or shade them).
- Make sure that you don't have other rows or columns with other data or formulas, you don't want Excel to read these as part of your dataset. Also, if you have hidden rows or columns, unhide them.

### How to do it:

- Highlight all the cells in the dataset, including the variable names.
- **WARNING:** Don't use the shortcut of just highlighting the columns because that will include all 65,536 worksheet rows in the named range.
  - Go to the INSERT menu, select NAME, and click on CREATE. By default, it will ask if you want to use the top row as names. Click OK.
  - Now you want to name the whole dataset, not just each column of data. Highlight all the cells in the dataset, including the names in Row 1.
  - Go to the INSERT menu, select NAME, and click on DEFINE.
  - In the box where the cursor is blinking, write the name of the dataset. I recommend that you use the name database because Excel expects this by default, so it will save you a little work later. Click OK.

### Other fun stuff:

The name box is on the far left of your formula bar underneath the font box. If you click on the arrow by the box, it will show you the named ranges that you've created. Click on a name and it will highlight the range (but not the cell with the variable name).

If you want to check the ranges associated with the names, go back to INSERT and NAME, then select DEFINE. As you select different names, you'll see the range at the bottom of the dialog box. Notice that the range includes the sheet name. *This really simplifies writing formulas!*

## Using formulas

---

### General points about formulas:

- Formulas start with an = sign followed by the name of the formula, then parentheses that contain the cell(s) that the formula will use. For example, =AVERAGE(D2:D346) will give you the average age of Needs Assessment respondents. NOTE: You always have to have complete sets of parentheses.
- Put the formula in the cell where you want the result of the calculation to appear.
- Excel doesn't care whether you use upper or lower case. After you enter the formula, it will return it in upper case.
- You can cut and paste or copy and paste formulas which makes it easy to replicate them for many cells. This by default, so it will save you a little work later. Click OK.
- If you want the reference cell to change as you copy the formula, (this is called a relative reference), just type the cell location (e.g., A10).
- If you want the formula to refer to a fixed location no matter where you copy it, use a dollar sign before the parts of the cell location you want to fix (e.g., \$A10 or A\$10 or \$A\$10).
- When you want to refer to cells in a different worksheet, you have to include in the formula the sheet name followed by an exclamation point and then the cell or cell ranges (e.g., data!A2:A346).
- If you've named your ranges, all you have to include is the name of the range (no sheet name, no cell ranges) and Excel automatically knows where to look. For example, to get average age of Needs Assessment respondents, type =AVERAGE(AGE). See I told you it would be a timesaver.

## Common formulas

---

There are LOTS of different formulas and permutations of formulas you can use. You can search under functions in the HELP feature for more.

- **Totals**  
You won't use this much for analyzing your evaluation data, but the formula is  
**=SUM(range)**
- **Averages**  
This returns the average or mean value for a group of data. Before you calculate an average, be sure it's something that makes sense for your variable. For example, average age makes sense, but does average race? The Needs Assessment data has six race/ethnicity categories, what does an average value of 1.64 mean?  
**=AVERAGE(range)**
- **Standard Deviation**  
This is a measure of spread around the mean. It gives people who are familiar with statistics a quick way to judge whether the data are clustered around the mean or spread out. For those of you who had statistics 10 years ago, this probably rings a bell, but you'd be hard-pressed to know what constitutes a narrow or wide distribution.  
**=STDEV(range)**
- **Other calculations**  
Assuming you already know this, but you can add, subtract, divide, or multiply cells. For example:
  - Add: **=(B2+D2)**
  - Subtract: **=(B2-D2)**
  - Divide: **=(B2/D2)**
  - Multiply: **=(B2\*D2)**

## Using pivot tables

---

Pivot tables allow you to look at the relationships between two or more variables. Here are some questions you can answer with the Needs Assessment data using pivot tables. Do men and women prioritize the APDP program differently? Does level of illness determine how likely people are to use case

management? Are there racial/ethnic differences in the prioritization of housing? Other programs refer to these analyses as crosstabs. In Excel, they are called pivot tables.

Most of the pivot tables that you will be using will rely on frequencies, in other words, counts and percentages of the number of cases that fall into a certain category. Because Excel is geared toward budgets and spreadsheets, the manuals and help section tend not to be very helpful when you want to find out how to do frequencies. Therefore, this manual focuses only on frequencies. Once you get frequencies under your belt, you'll be able to explore the different pivot table options on your own to find other uses that may help you with other parts of your job.

### **Things to do before you begin:**

- Make sure you've removed all filters.
- Name all the ranges that you'll be using. In general, this means name everything you can name.
- Display the pivot table toolbar: go to the VIEW menu, click on TOOLBARS, then select PIVOT TABLE (in version 5.0, you will select QUERY AND PIVOT).

### **Pivot Table Wizard**

Pivot Table Wizard will take you through the steps necessary to create a pivot table. As with most Excel wizard features, it is easy to use after you know how to use it and the help feature is often less than helpful.

## **Pivot table steps**

---

### **Step 1**

Place your cursor anywhere in your data. Start the PT wizard by either clicking the PT button on the PT toolbar or going to the data menu and selecting Pivot Table Report. This opens the Step 1 dialog box which asks for the location of your data. Your data are in an Excel spreadsheet, so select Microsoft Excel list or database and then click on NEXT.

### **Step 2**

Step 2 asks for the ranges of the data in the spreadsheet. By default, it assumes that you've named your database database. If you didn't use the name database put the name that you used into the box. If you haven't already named your data, you can select the range that you want using these steps. First, shrink the dialog box by clicking on the box with the red arrow next to the range box. Now, highlight the entire dataset and click on the dialog box's expand button. (NOTE: If you haven't named the database, highlighting the data before you go into wizard won't work. You have to highlight the data while in wizard.) You've now inserted the appropriate range in the dialog box. Click on NEXT.

### **Step 3**

Step 3 shows a representation of the table with your data choices as buttons on the right. On this step, you drag the buttons for the data you want to analyze into the row, column, and data areas. At the very least, you must drag one button into the row or column area and one button into the data area. See the count, don't sum section below for more details about how you want Excel to analyze your data. When you have dragged all the buttons to where they belong, click on NEXT.

### **Step 4**

Step 4 asks where you want the pivot table to go. If you left the cursor somewhere in the middle of your dataset, it will default to asking if you want the table in a new worksheet. If the cursor is already in a new sheet, it will default to asking if you want it in the existing sheet. I recommend that you don't put any tables in the sheet that has your data. Select where you want the table to go and then click FINISH.

## Do men and women prioritize APDP programs differently?

---

### Example:

I want to use the Needs Assessment data to look at whether men and women prioritize the APDP program differently.

1. At Step 4, I drag the sex button into the column area and the priordrugs button into the row area (or vice versa).
2. Because I am interested in the APDP priority information broken down by sex, I go back to the priordrugs button on the right side and this time I drag it into the data area. Notice that Excel changes the name of the button to sum of priordrugs.
3. I can drag any button into the data area as many times as I want.
4. In anticipation of a later step, I'll drag the priordrugs button into the data field for a second time. Excel renames this button sum of priordrugs2.
5. If I were to finish the wizard steps right now, Excel would sum all of the numbers in the cells associated with each of the categories I've designated for the table.

But I don't want to do that, so read the next section...

### Count, don't sum

Because Microsoft created Excel to use for budgets and spreadsheets, most of the defaults are geared in that direction. For pivot tables, this means that the program assumes that you want to sum the numbers in the columns of data (if you have text in the columns, then it will automatically count how often the text occurs). If you want to do anything else, like count them, you have to take a few extra steps to get there. Try clicking the FINISH button after the last example. Do the results make sense for creating a frequency table? Nope.

To create frequencies, you want to count the number of cells in a column that contain a specific number. For example, if you want to know how many women are in the Needs Assessment dataset and the code for female is 2, you want to count how many twos there are in sex column. Likewise if you want to count how many women prioritized the APDP program (i.e., sex=2 and priordrugs=1), you want to count how many rows contain both a 2 in the sex column and a 1 in the priordrugs column.

To change the sum function to a count function, double-click on the appropriate button in the data field. Highlight COUNT and click OK. Continue through the remaining wizard steps.

### Example (continued):

1. Double-click on sum of priordrugs and that opens a dialog box which shows the source field (i.e., where the data are coming from for the calculation) and the options for using different calculations.
2. Using the cursor, highlight COUNT, click OK, and finish using wizard.

Now when I look at the pivot table, it shows how many cells it counted with each of the possible combinations of **sex** and **priordrugs**.

### Showing percentages

Now you know that 6 of the 38 women and 91 of the 297 men in the Needs Assessment dataset prioritized the APDP program. But what are those figures in terms of percentages?

To have Excel calculate the percentages:

1. Go back into Step 3 of the Pivot Table Wizard. You can get there by placing your cursor in any cell of the existing pivot table, and clicking on the pivot table symbol on the pivot table toolbar.
2. Double-click on the second data field button, sum of priordrugs2.
3. Highlight COUNT and then click on the OPTIONS button.
4. In the SHOW DATA window, click on the down arrow, and highlight % OF COLUMN. (NOTE: If you put sex in the rows and priordrugs in the columns, then highlight % OF ROW.)

5. Click OK. You can change the number of decimal places in the percentages by clicking the NUMBER button before you click on OK. Then you select the number of decimals you want and move back through the boxes by clicking OK.

### Averages

You might not always want to calculate frequencies. Say you have a scale that ranges from 1 to 10 and you want to compare the average scale score for white people and people of color.

To calculate averages in your pivot table, place your scale button in the data field, double-click on it, and choose AVERAGE instead of count.

### Changing item names

Now you look at your table and instead of seeing labels like "male" and "female," you see the numbers that represent categories like male and female in the dataset. How do you change those labels?

Select the cell that you want to change (don't worry when you see the whole column or row highlighted, you can't change the data inside the pivot table without changing the original dataset). Type in the label you want. If you want to change the titles of the calculations (e.g., count of priordrugs2), click on one of the cells with that name (all associated rows or columns will be highlighted), and type in % of column or whatever name you want.

### Hiding data

If you look at the Needs Assessment table we've created, there are more rows and columns than just male, female, yes, and no. You may want to hide some of the rows and therefore eliminate them from the calculations. Take the example of the sex and prior drugs table. On the sex variable, the codes 3 and 4 refer to transgendered people (Male to Female and Female to Male, respectively). Only four people fit into these categories. Plus there is a column with a period—this represents missing data. If you only want to look at men and women, while omitting transgendered people and missing data, you can hide this data.

Go back into Step 3 of the wizard. Double-click on the button for the variable with the categories you want to hide. At the bottom of the dialogue box, you'll see a window that says **Hide items**. Highlight those categories that you want to hide by clicking on each one. Click on OK and then FINISH.

### General comment about missing data

If you have just a handful of cases with missing data (5 percent or less), then it's ok to take them out of the calculations. If you have more than a handful, keep them in the table so readers can see if these missing data cases "look" different than cases on which you could collect data.

## Recoding Variables with VLOOKUP

---

### General information

Sometimes you collect information in one form and then you need to recode it into another form before you can use it for your analyses. For example, the Needs Assessment data contains information about people's age, but you might want to group people into age categories in order to create a pivot table. Or say you've got information about race/ethnicity, but you have so few people of color in each category that you want to compare white people to people of color as a group.

Statistical programs like SPSS make this easy. Unfortunately, the procedure in Excel is more cumbersome, but it is still very doable.

### VLOOKUP

The function that you will use is called VLOOKUP. Essentially, the program compares the values of the variable that you want to recode to a list of new values that you specify and then returns a new value based on this list. It sounds complicated, but if you follow these step-by-step instructions, it is actually pretty easy. Be forewarned, this is a pared down explanation of VLOOKUP. If you want to do anything more complicated or substantially different than what I describe here, find a manual and read about the function in more depth.



## Syntax

"Syntax" is a generic representation of the structure of a formula. When you use the formula, you substitute actual numbers and ranges into the formula. Below, I show the syntax for VLOOKUP and give descriptions of each part of the syntax. This is easiest to follow using an example, so let's use the example of recoding age into categories.

The general form of the function is:

VLOOKUP(lookup\_value,table\_array,col\_index\_num,range\_looku

## Definitions

- The **look\_up value** refers to the cell with the old value of age that you want to change.
- The **table\_array** specifies the entire range of numbers that contains old values matched with new values. You will create this two-column array on a separate sheet in your workbook.
- The **col\_index\_num** tells Excel which column in the array contains the recoded values. You'll put the recoded values in column 2.
- The **range\_lookup** argument tells Excel how to compare the look\_up value with the first column in the table\_array. To have Excel look for an exact match, you'll enter FALSE.

## Example: Recoding age into categories

The Needs Assessment data asked for people's age rather than asking them to check the age category into which they fall. I want to create a new variable called AGECAT, with the following values:

1 = under 20

2 = 20-29

3 = 30-39

4 = 40 or older

**Step 1:** Open a new worksheet and name it VLOOKUP\_age. I use this name so I easily know the location of my recode values.

**Step 2:** In the first column (column A), I put all possible ages that might occur in the dataset. The oldest respondent is 68, so I'll fill the column with numbers from 1 to 68. Excel has an AUTOFILL feature that makes this easy. I put the first three numbers of the series (1, 2, 3) into the first three cells of column A and I highlight those cells. Then, I position my cursor on the little black box on the lower right corner of the highlighted area, the cursor changes to a little black cross, and I click and drag down to the 68th row. Excel automatically fills the empty cells with numbers that continue the pattern of the first three cells I highlighted. (You can do this with almost any series, even things like months, as long as you give Excel enough information to recognize the pattern.)

**Step 3:** In the second column (column B), place the appropriate recode value next to the original age value, using the coding scheme above. Excel makes this easy using the same procedure as with AUTOFILL. Put a 1 in the cell next to 1, highlight that cell, click and drag the little black box down to the cell next to 19, and all highlighted cells fill with ones. Repeat the procedure with the 2s, 3s, and 4s.

**Step 4:** Name this range (also called the table\_array). Highlight all the data in both columns. Go to the INSERT menu, select NAME, and then DEFINE. The highlighted range will show in the window at the bottom of the box. In the top window where the cursor is blinking write agerecode.

**Step 5:** Now go back to the worksheet that contains your data. Insert a column to the right of the original age column (this will automatically include the new variable in the database range without you having to



redefine that range). At the top of the new column, write the name of the new variable, AGE CAT (or whatever).

**Step 6:** Now you will enter the vlookup formula that will do the recode. Place your cursor in the cell immediately below the variable name (E2). Type the following text and then hit ENTER:

**=vlookup(D2,agerecode,2,FALSE).**

When you hit ENTER, Excel will return the new value.

**Step 7:** Copy and paste this command to the remaining cells in the column.

What does that formula mean? You've told Excel to use the VLOOKUP function to exactly match the value in cell next door (in the age column) with the numbers in the first column of the agerecode array and return the corresponding number in column 2 of that array. The value in cell D2 is 37, Excel looks at the agerecode array finds 37 in the first column and sees that it is paired with a 3 in the second column, so it returns a 3 to the cell that contains the formula (cell E2).

## Creating charts

---

### Why use charts?

Charts, like other visual aids, enhance your presentation because they can convey a lot of information at a glance. Relationships buried in text can pop out when displayed in a chart. Once you have pivot tables down pat, you can create tables that readily translate into a chart. There are a gazillion variations on the charts you can make. I'll go over one and then, as you get more comfortable, you can play around and explore your options.

### Chart Wizard

Chart Wizard takes you through the steps necessary to create a chart.

First: Create a pivot table with either percentages or counts. If you include both, Excel won't know which numbers to use. Highlight the table (you can select the whole table by clicking on its upper left corner). Click on the Chart Wizard button on the toolbar.

Step 1: Select the type of chart you want. Each type (column, bar, pie, etc.) has several variations, so you'll need to select the variation you want. Because I'm looking at the relationship between two variables, I'll choose a column chart with two bars per category (as opposed to the categories being stacked on each other). Click on NEXT.

Step 2: Excel will show you what the chart is going to look like and where it's getting the data from. Click on the tab that says Data Series. In the window on the left you can add and remove data series. Highlight GRAND TOTAL and click on REMOVE. The grand total column equals 100 for each category, so it's not adding any information to the table. Click on NEXT.

Step 3: This is the step where you tweak the formatting. You can add titles, labels, gridlines, etc. Experiment to see how you like your chart to look. Click on NEXT.

Step 4: Tell Excel whether you want your chart in the same sheet as the table or in a separate worksheet. Click on FINISH.

If you need to make changes to your chart, click on the chart and then click on the Chart Wizard button. This will put you back at Step 1. You can also customize the chart by double-clicking on various parts to bring up dialog boxes with more options. You can change the size of the chart by clicking on it and then dragging on the lower right corner of the box that contains the chart.